The Simulated Online EM Algorithm for Latent Factor Models

D. Rohde[*], O. Cappé[†] and O. Dikmen[**]

[*] *Instituto de Matemática, UFRJ, Rio de Janeiro, Brazil*
[†] *CNRS & LTCI Télecom ParisTech, Paris, France*
[**] *School of Science and Technology, Aalto University, Helsinki, Finland*

Abstract. The estimation of latent factor models are treated in an integrated maximum likelihood context where one parameter is marginalized and another is estimated. An extension to the online Expectation Maximization (EM) algorithm is employed the simulated online Expectation Maximization algorithm. Both these algorithms apply to exponential family models, but the simulated version of the algorithm can make use of Monte Carlo simulation to compute the stochastic E-steps while maintaining the convergence properties of the original online EM algorithm. A class of important latent factor models are identified that can be expressed in complete data exponential family form, the algorithm is applied to one of these models Itakura-Saito Non-negative Matrix Factorisation. An additional parameter is introduced into this model and it is conjectured if this is set to a high value the posterior variance of the parameters is reduced and estimation becomes easier. Simulations are provided that support this conjecture, although online estimation for models with even a modest number of components continues to be hampered by the presence of local minima.

Keywords: MCMC, semi-Bayesian, online estimation
PACS: 02.50.Ng,02.50.Tt

INTRODUCTION

The online EM algorithm is a variant of the EM algorithm that preserves many of the appealing features of the original EM algorithm in an online setting [4, 2], it is of particular relevance when the data being analysed is large and perhaps growing and the conventional or batch EM algorithm may be too slow. The online EM algorithm is applicable to complete data exponential family models in which there is the availability of an analytical E-step which computes the expectation of the complete data sufficient statistics. The simulated online EM algorithm studied here allows estimation when an analytical E-step is not available using Monte Carlo methods, this algorithm also converges albeit with a higher variance[13].

In this paper we study the applicability of this algorithm to latent factor models. Latent factor models find applications in many fields, non-negative matrix factorization models are a prominent example which are widely applied to decompose images or audio spectra into components, an advantage of a non-negative constraint is that the decomposition is often readily interpretable [11]. Other prominent applications of latent factor models include topic modelling where documents in a corpus are clustered into topics where any document may have multiple topics [1] and collaborative filtering applied to recommender systems [14].

As the online EM algorithm is based on the exponential family of distributions, a

modelling framework based upon the exponential family is required. Such a framework is developed for latent factor models in [10] which show that three important matrix factorization models can be put in exponential family form by augmenting with appropriate latent data. These authors use this framework to demonstrated how a Gibbs sampler can be constructed which exploit the exponential family in two distinct places. In our work here, the first use of the exponential form is necessary for allowing estimation in an online EM framework. The second is useful although not necessary for using the Gibbs sampler as the sampling procedure. We also consider the Metropolis-Hastings sampler.

In latent factor models it is possible to estimate all parameters including the latent factors and a large literature is devoted to this approach much of it inspired by the seminal work in [11]. Although these methods which estimate all parameters are widely deployed with some success this approach does have some problems in particular the the number of parameters grows with the number of data points, which makes analysis of convergence difficult and requires the employment of heuristics to apply and validate the model on unseen data. A growing literature shows that improved statistical performance is obtained by operating in a semi-Bayesian framework where the latent factors are marginalized out and other parameters are estimated [1, 6, 7]. The online EM algorithm developed here operates in such a semi-Bayesian framework. The E-step must marginalize both over the introduced latent data as well as the latent factor models. This is in contrast to typical EM algorithm approaches to matrix factorisation that simply marginalize over introduced latent data and estimate all other parameters including the latent factors e.g. the SAGE algorithm in [9].

In this contribution we apply one model from [10], Itakura-Saito Non-negative Matrix Factorisation or IS-NMF. The simulated online EM algorithm has already been applied to this model in [3] (in French). Our contribution here is to show that with a trivial modification to the model and applying a slightly modified computational procedure we obtain an algorithm with lower posterior variance. Simulations demonstrate that estimates converge to a local minima much more quickly and with less noise. For example images that are learned are cleaner in appearance if M is high compared to M low. Unfortunately in our simulation studies the local minima is rarely a global minima.

SIMULATED ONLINE EM ALGORITHM

The online em algorithm is applicable to a latent data problem where (x,y) are the latent and real data of a complete data exponential family model i.e. where the model has the following exponential family form

$$\mathrm{P}(x,y|\theta) = h(x,y)\exp\{\phi'(\theta)s(x,y) - A(\theta)\}$$

where y is observed, $x = [H\ C]$ is latent, $\phi(\cdot)$ maps the parameter θ to the natural parameters and $s(\cdot)$ maps (x,y) to the sufficient statistic and $A(\cdot)$ is the log partition function. The online EM algorithm then operates by alternating the following steps.

The stochastic E-step

$$S_n = (1-\gamma_n)S_{n-1} + \gamma_n \mathrm{E}_{\bar{\theta}(S_{n-1})}[s(X_n,Y_n)|Y_n]$$

The stochastic M-step

$$\theta_n = \bar{\theta}(S_n)$$

Where S_n is the current estimate of the complete data sufficient statistics for the data set $Y_1, \ldots Y_n$, as it is an online algorithm n indexes both the amount of data observed so far and the number of iterations; θ_n similarly denotes the parameter estimate after n iterations, γ_n is a decaying sequence satisfying $\sum_{n=0}^{\infty} \gamma_n = \infty$ and $\sum_{n=0}^{\infty} \gamma_n^2 < \infty$, empirically a good choice is $\gamma_n = n^{-0.6}$.

The algorithm and proof of convergence is described in detail in [4, 2]. The simulated online EM algorithm replaces $E_{\bar{\theta}(S_{n-1})}[s(X_n, Y_n)|Y_n]$ with a Monte Carlo simulation with the same expectation. This algorithm also converges albeit with higher variance [13]. A straightforward means to produce samples with this expectation is to take R samples of $\tilde{X}_n^r \sim P(X_n|Y_n, \theta)$ and then to compute $\frac{1}{R}\sum_{r=1}^{R} s(\tilde{X}_n^r, Y_n)$. Alternatively Rao-Blackwelization may be available to reduce the variance, in particular it maybe convenient to sample only a sub-component of $\tilde{X_n}$ i.e. where $X_n = [H_n\, C_n]$ sample only $\tilde{H_n}|\theta, Y_n$ and then compute $E_{\bar{\theta}(S_{n-1})}[s(X_n, Y_n)|\tilde{H_n}]$ which has reduced variance, and avoids directly sampling C_n.

In practice instead of generating independent samples from H_n or X_n a Markov chain Monte Carlo algorithm is used with a carefully chosen initial state and long burn in to reduce the bias. In the discussion here we assume that the bias can safely be neglected.

LATENT FACTOR MODELS

The models we consider have the following structure: we model the observed data Y which is an $F \times N$ matrix, with a parametric model such that there is a parameter θ_f for every one of the F columns of the matrix and such that there is a parameter H_n for ever one of the N rows of the matrix. We can therefore model the data $P(Y_{f,n}|\theta_f, H_n)$, and conditional on the parameters, the elements of Y are independent. In order to obtain exponential family form it is required to augment each $Y_{f,n}$ with latent data C.

A fully Bayesian treatment of these models can be approximated with the following Gibbs sampling algorithm [10] which can be applied efficiently to models that make use of the exponential family of distribution in two distinct places.

1. $P(\theta|Y, C, H) \propto P(\theta) \prod_{n=1..N} P(Y_n, C_n, H_n|\theta)$
2. $P(H|Y, C, \theta) \propto P(H) \prod_{f=1}^{F} P(Y_f, C_f, \theta_f|H)$
3. $P(C|Y, \theta, H)$

In order for step 1 to be (easily) sampled using a Gibbs sampler $P(Y_n, C_n, H_n|\theta)$ must be in exponential family form where Y, C, H are treated as data observed (Y) or latent (C, H) and θ is the parameter. Similarly in order for step 2 to be (easily) sampled using a Gibbs sampler $P(Y_f, C_f, \theta_f|H)$ must also be in exponential family form where Y, C, θ are treated as being data real (Y) and latent (C, θ) and H is the parameter. In this framework the exponential family model is being used in two distinct places in the first case H is seen as latent data and θ as a parameter and in the second H is seen as a parameter and θ as latent data. In a fully Bayesian treatment this double use of exponential family

models is advantageous in allowing the use of Gibbs sampling in turn sampling θ and then H. The use of two exponential family models are similarly useful in our semi-Bayesian framework because it allows estimation of θ in an online EM setting and Gibbs sampling can again be used for the integration of C,H (required for computing $\mathrm{E}[s(Y,C,H)|Y]$ for the online EM stochastic E-step), although other samplers such as random walk Metropolis-Hastings may also be used.

Itakura-Saito NMF

Lazy completion

The IS NMF model has the following form

$$V_{f,n} = Y_{f,n}^2 \sim \Gamma(M/2, \frac{2}{M}\sum_{k=1}^{K}\theta_{f,k}H_{k,n}) \tag{1}$$

where $V_{f,n}$ is observed directly and $Y_{f,n}$ is its square. This equation refers to individual elements of the $V \times F$ matrices and an individual element of the $F \times K$ matrix θ. The following parameterization of the Gamma distribution is used

$$f(x;\nu,\lambda) = x^{\nu-1}\frac{e^{-x/\lambda}}{\theta^{\nu}\Gamma(\nu)}.$$

which has expectation $\nu\lambda$, so that we have $\mathrm{E}[V_{f,n}|\theta_{f,k},H_{k,n}] = \sum_{k=1}^{K}\theta_{f,k}H_{k,n}$, giving the justification of the name probabilistic matrix factorization.

Full completion, exponential family form

For every observed $Y_{f,n}$, a $K \times M$ matrix of latent data is introduced, the latent data C is therefore indexed $C_{f,k,m,n}$. Conditional on θ and H the elements of $C_{f,k,1,n},...,C_{f,k,M,n}$ are independent and have the following distribution

$$C_{f,k,1,n},...,C_{f,k,M,n}|\theta_{f,k}H_{k,n} \sim \mathcal{N}(0, \frac{\theta_{f,k}H_{k,n}}{M}).$$

which means we can say

$$\sum_{m=1}^{M} C_{f,k,m,n}|\theta_{f,k}H_{k,n} \sim \mathcal{N}(0, \frac{\sum_k \theta_{k,n}H_{k,n}}{M}).$$

If V is the matrix of observations and its elements are defined as $V_{f,n} = Y_{f,n}^2$ where

$$Y_{f,n}^2 = \sum_{k=1}^{K}(\sum_{m=1}^{M} C_{f,k,m,n})^2,$$

by the definition of the χ^2_M distribution we can say that

$$\frac{M|Y_{f,n}|^2}{\sum_k \theta_{k,n} H_{k,n}} \sim \chi^2_M().$$

Using the equivalence between gamma distribution and χ^2 distributions and the scaling property of the gamma distribution we recover the matrix factorization model given in Equation 1. Due to the semi-Bayesian treatment a prior is needed on H_n, for computational convenience an independent inverse gamma prior is put on each $H_{k,n}$. The full semi-Bayesian expression for the model then becomes

$$\begin{aligned}
&P_\theta(C_{1,1,1,n},...,C_{F,K,M,n},H_{1,n},...,H_{K,n}) \\
&= (\prod_{k=1}^{K}(H_{k,n})^{-\alpha-1}\exp\{-\frac{\beta}{H_{k,n}}\}) \\
&\prod_{f=1}^{F}\prod_{m=1}^{M}\frac{1}{\sqrt{2\pi}}\exp\{\frac{-C^2_{f,k,m,n}M}{2\theta_{f,k}H_{k,n}} - \log(\frac{\theta_{f,k}H_{k,n}}{M})\} \\
&= (\prod_{k=1}^{K}(H_{k,n})^{-\alpha-1}\exp\{-\frac{\beta}{H_{k,n}}\})(2\pi)^{-\frac{FM}{2}} \\
&\exp\{\sum_{f=1}^{F}\sum_{m=1}^{M}\frac{-C^2_{f,k,m,n}M}{2\theta_{f,k}H_{k,n}} - \log(\frac{\theta_{f,k}H_{k,n}}{M})\}
\end{aligned}$$

which is in exponential family form with sufficient statistics:

$$\begin{pmatrix} \frac{C^2_{1,1,1,n}}{H_{1,n}M} \\ . \\ \frac{C^2_{F,1,1,n}}{H_{1,n}M} \end{pmatrix},...,\begin{pmatrix} \frac{C^2_{1,K,M,n}}{H_{K,n}M} \\ . \\ \frac{C^2_{F,K,M,n}}{H_{K,n}M} \end{pmatrix}$$

which is $K \times M$ vectors of length F.

Stochastic E-step

Practical implementations of the simulated online EM algorithm proceeds by using MCMC samples of C,H in order to approximate the E-step. One possible sampler is the block Gibbs sampler which exploits the exponential family form given above and the derivation for which is given in [10]. These authors focus upon cases where $M = 1$ and $M = 2$, this means that the size of C at $F \times K \times M$ is manageable, but it is clear that for large M then C becomes a large array requiring significant memory. This model is often applied to signal processing problems, $M = 1$ is therefore the simplest form of the model the so called real IS NMF model as it is easy to formulate with a real normal distribution. Similarly $M = 2$ is referred to as the complex model and can be formulated using the spherical complex normal distribution. In signal processing this is

an intuitively attractive model because spectra are complex. The properties of the model with $M > 2$ are unexplored in the current literature.

In this paper we propose using M large but avoid the extravagant representation of C by using the Metropolis algorithm on the lazy form of the model. The exponential family form is still required for the online EM algorithm however the expected sufficient statistics $\mathrm{E}_\theta[s(X,Y)|Y,H]$ can be computed either with a Gibbs sampler using the exponential family model or another MCMC algorithm using the lazy form of the model. The extravagant representation on C can again be avoided by making use of a Rao-Blackewelized expression of the sufficient statistics.

The Metropolis-Hastings algorithm [12] constructs a Markov chain with move proposals which are taken with an acceptance probability. As the parameters for this model are constrained to be non-negative a multiplicative log normal proposal is used instead of the normal additive Gaussian random walk. So the proposal $Q()$ has the form $Q(H^*|H) \sim H \times \log N(\mu, \sigma^2)$. This results in the following factor modifying the usual random walk Metropolis acceptance ratio

$$\min\left(\frac{Q(H,H^*)P(H^*|Y_n)}{Q(H^*,H)P(H|Y_n)},1\right) = \min\left(\frac{H^*\,P(H^*|Y_n)}{H\,P(H|Y_n)},1\right).$$

Rao-Blackwelized formula for computing Expectation of sufficient statistics

As $C_{f,1,m,n},...,C_{f,K,m,n}|H_n,\theta_f \sim \mathscr{N}()$ then $C_{f,1,m,n},...,C_{f,K,m,n},\sum_{k=1}^K C_{f,k,m,n}|H_n,\theta_f$ is a reduced rank normal distribution and its form can be computed by using the Affine transform rule for multivariate normal distributions. This results in

$$\begin{pmatrix} C_{f,1,m,n} \\ . \\ . \\ C_{f,K,m,n} \\ \sum_{k=1}^K C_{f,k,m,n} \end{pmatrix} |\theta,H \sim$$
$$\mathscr{N}\left(\begin{pmatrix} 0 \\ . \\ . \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{\theta_{f,1}H_{1,n}}{M} & 0 & . & \frac{\theta_{f,1}H_{1,n}}{M} \\ 0 & \frac{\theta_{f,2}H_{2,n}}{M} & . & \frac{\theta_{f,2}H_{2,n}}{M} \\ . & . & . & . \\ \frac{\theta_{f,1}H_{1,n}}{M} & \frac{\theta_{f,2}H_{2,n}}{M} & . & \sum_{k'=1}^K \frac{\theta_{f,k'}H_{k',n}}{M} \end{pmatrix}\right)$$

From this we can see that

$$C_{f,k,m,n}|\sum_{k=1}^K C_{f,k,m,n},\theta,H$$
$$\sim \mathscr{N}(\frac{\theta_{f,k}H_{1,n}}{\sum_{k'=1}^K \theta_{f,k'}H_{k'n}}(\sum_{k=1}^K C_{f,k,m,n}), \frac{\theta_{f,k}H_{k,n}}{M}(1-\frac{\theta_{k,n}H_{k,n}}{\sum_{k'1}\theta_{f,k'}H_{k',n}})$$

By using the identity that $\mathrm{E}[X^2] = \mathrm{E}[X]^2 + \mathrm{Var}[X]$, we can conclude that

$$\mathrm{E}_\theta[\sum_{m=1}^{M} C_{f,k,m,n}^2 | \sum_{k=1}^{K} C_{f,k,1,n}, ..., \sum_{k=1}^{K} C_{f,k,M,n}, H_{k,n}]$$
$$= \left(\frac{\theta_{f,k} H_{k,n}}{\sum_{k'=1}^{K} \theta_{f,k'} H_{k'n}} \right)^2 (\sum_{m=1}^{M} (\sum_{k=1}^{K} C_{f,k,m,n})^2) + \theta_{f,k} H_{k,n} (1 - \frac{\theta_{k,n} H_{k,n}}{\sum_{k'=1} \theta_{f,k'} H_{k',n}})$$

or more usefully

$$\mathrm{E}_\theta[\sum_{m=1}^{M} C_{f,k,m,n}^2 | Y_{f,n}, H_{k,n}]$$
$$= \left(\frac{\theta_{f,k} H_{k,n}}{\sum_{k'=1}^{K} \theta_{f,k'} H_{k'n}} \right)^2 Y_{f,n}^2 + \theta_{f,k} H_{k,n} (1 - \frac{\theta_{k,n} H_{k,n}}{\sum_{k'=1} \theta_{f,k'} H_{k',n}})$$

Finally this allows the following Rao-Blackwelized expression for the expected sufficient statistics, $S_{f,k,n} = \mathrm{E}_{\theta_{n-1}}[\frac{\sum_{m=1}^{M} C_{f,k,m,n}^2}{H_{k,n}} | Y_{f,n}, H_{k,n}] = \frac{\mathrm{E}_{\theta_{n-1}}[\sum_{m=1}^{M} C_{f,k,m,n}^2 | Y_{f,n}, H_{k,n}]}{H_{k,n}}$.

Stochastic M step

In the online EM setting stochastic approximation will approximate the expected sufficient statistics $\tilde{S}_{f,k,n} \approx \frac{1}{n} \sum_{n'=1}^{n} \sum_{m=1}^{M} \frac{C_{f,k,m,n}^2}{H_{k,n} M}$, and will equal this value as $n \to \infty$, assuming the approximation is good we obtain

$$\theta_{f,k} | C_{f,k,1,1}, ..., C_{f,k,M,n} \sim \mathcal{IG}(\alpha + n, \beta + n S_{f,k,n}).$$

$$\bar{\theta}(S_{f,k,n}) = \frac{\beta + n S_{f,k}}{\alpha + n + 1} = \frac{\beta/n + S_{f,k}}{\frac{\alpha+1}{n} + 1}$$

The effect of the M parameter

If we consider the model in lazy form then $V_{f,n} \sim \Gamma(M/2, \frac{2}{M}\lambda)$ where $\lambda = \sum_{k=1}^{K} \theta_{f,k} H_{k,n})$, we find that $\mathrm{E}_\theta[V_{f,n} | H_n] = \beta$ and the variance is given by $\mathrm{Var}_\theta[V_{f,n} | H_n] = 2\frac{\lambda^2}{M}$, it is apparent that M controls the variance of the predictive distribution. It is conjectured that using M large may increase the identifiability of the parameters and make estimation easier.

A simple way to investigate if the conjecture is correct is to draw samples of data from the model with different values of M. In Fig 1 data is compared to the noiseless signal

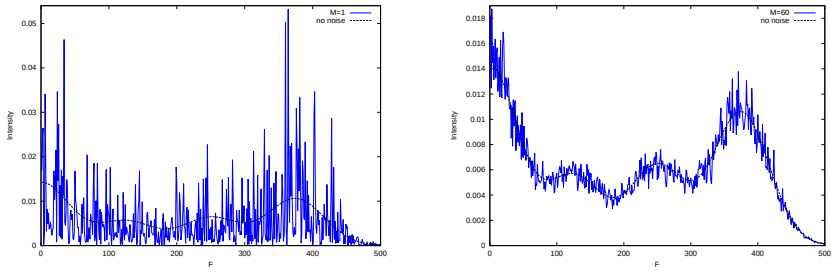

FIGURE 1. Data simulated from two models that differ only in the parameter M, left $M = 1$ right $M = 60$.

i.e. θH_n with samples of data from the model with $M = 1$ and $M = 60$, it is very clear that much more noise is present in the $M = 1$ model.

To further consider this point we can also investigate the posterior of $H_{n,k}$ conditional on the same value of θ and the same data. This is presented in Figure 2, again it is evident that the posterior is much narrower for $M = 60$ than for $M = 1$ or $M = 2$.

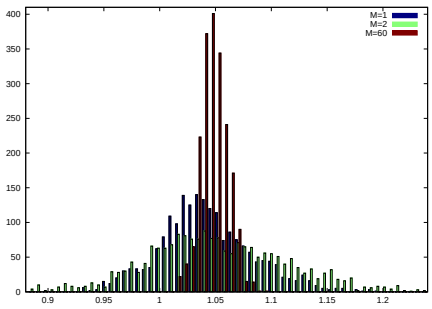

FIGURE 2. Comparison of posteriors for $P(H_{k,n}|Y_n,\theta,M = 1)$, $P(H_{k,n}|Y_n,\theta,M = 1)$ and $P(H_{k,n}|Y_n,\theta,M = 60)$.

Finally we consider the actual operation of the simulated online EM algorithm. As a test problem we consider the 'swimmer' dataset [8] which consists of a swimmer with 4 limbs each of which has 4 positions. A correct separation produced by a non-negative matrix factorization algorithm consists of separating each of the limbs into 16 components, a further element, the body, is not identifiable and may be shared between the components.

The simulated online EM algorithm was successfully applied to this dataset in [13] using the Latent Dirichlet Allocation model. The IS NMF model with $M = 1$ was also successfully used with this model using a batch variational Bayes approach in [7]. These authors experience difficulties with the EM algorithm getting caught in local minima, which they deal with by using two strategies, by doing multiple runs, picking the run with the largest marginal likelihood and by using an annealing procedure. Even when employing these procedures the estimates of the limbs of the swimmer are noisy, see Figure 2 in [7].

The overall result from applying the online EM algorithm to this problem with $M = 60$, is that the estimates of the limbs are no longer noisy as they were when applied

to the model with $M = 1$, however the problem of falling into local minima i.e. failing to separate the limbs is acute. Moreover the heuristics for avoiding local minima such as annealing or comparing marginal likelihoods of multiple runs used in [7] are not easily available in the online context.

The swimmer dataset was generated in the following way. If a pixel was considered to be black then it was given the low value of 0.01, if it was white it was given a value of 10, the 256 images (32 pixels square) were then flattened into vectors of length 1024, data was then generated using these templates from the model with $M = 60$. It should be noted that this results in much less noisy data than generating with $M = 1$ or $M = 2$. The algorithm was run with $K = 32$, when the ground truth is $K = 16$. This allows us to demonstrate the automatic order selection property observed in [7] that excessive capacity in θ automatically gets removed by estimating these components as being near zero. The algorithm was run for 20×256 iterations i.e. for 20 (noisy) repetitions of the 256 sized dataset, each stochastic E-step involved a burn in of 1400 iterations and averaging over 100 samples. The estimates of the 16 highest values of θ are shown in Figure 3. Unfortunately the algorithm seems to inevitably get caught in local minima, several limbs are not correctly separated. On the other hand the estimates of the limbs are much less noisy than cases that are applied with $M = 1$ or $M = 2$. The property of automatic order selection, is also evident here where unused components are estimated to be zero. This is one of the advantages noted in [7] of marginalizing H.

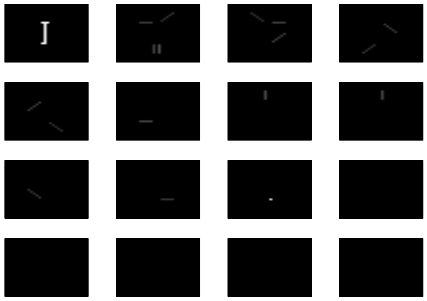

FIGURE 3. The estimates of θ for applying IS NMF on the swimmer problem with $M = 60$.

DISCUSSION

The simulated online EM algorithm is shown to be applicable to latent factor models presented in [10].

This work considered an elaboration of the IS NMF model presented in [10] which results in a matrix factorization model with an additional parameter M, some preliminary simulations support our conjecture that setting M to be high reduces the posterior variance of the parameters and results in a model for which parameter estimation is easier. In addition we also observe automatic order selection where unneeded components of θ are set to zero.

Unfortunately setting M high does not reduce the problem of local minima which appear to be a serious difficulty for the IS NMF model. Moreover, it is unclear how strategies used in a batch setting for avoiding local minima such as annealing or com-

paring the marginal likelihood of multiple runs [7] can be employed in an online setting. Annealing is difficult because the online EM employs stochastic approximation not for optimisation, but rather using the Robbins Monro stochastic approximation method for finding the root of an equation. Equally comparing multiple runs requires a computation of the marginal likelihood, it is difficult to take on this computation without employing a completely different computational procedure such as Chibb's method [5] for comparing the output of different algorithm runs. Alternatively the variational Bayes framework adopted in [10] could be used in order to optimise a lower bound on the integrated likelihood.

ACKNOWLEDGMENTS

We would like to thank Cedric Févotte for interesting discussions. David Rohde was partially supported by the Programa Professor Visitante do Exterior from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

REFERENCES

1. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *JMLR*, 3:2003, 2003.
2. O. Cappe. Online Expectation-Maximization. In K Mengersen, M. Titterington, and C. P. Robert, editors, *Mixtures*. Wiley, 2011. to appear.
3. O. Cappé, C. Févotte, and D. Rohde. Algorithme em en ligne simulè pour la factorisation nonnégative probabiliste. In *In Colloque du GRETSI*, Bordeaux, France, September 2011.
4. O. Cappe and E. Moulines. On-line expectation-maximization algorithm for latent data models. *J. Roy. Statist. Soc. B*, 71(3):593–613, 2009.
5. Siddhartha Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313, 1995.
6. O. Dikmen and C. Fevotte. Maximum marginal likelihood estimation for nonnegative dictionary learning. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011.
7. O. Dikmen and C. Févotte. Nonnegative dictionary learning in the exponential noise model for adaptive music signal representation. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24 (NIPS)*, pages 2267–2275. MIT Press, Granada, Spain, Dec. 2011.
8. D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In S. Thrun, L. Saul, and B. Schölkopf, editors, *NIPS 16*. MIT Press, Cambridge, MA, 2004.
9. C. Fevotte, N. Bertin, and J. L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. with application to music analysis. *Neural Comput.*, 21(3):793–830, Mar. 2009.
10. C. Fevotte and A. T. Cemgil. Nonnegative matrix factorisations as probabilistic inference in composite models. In *Proc. 17th European Signal Processing Conference (EUSIPCO)*, pages 1913–1917, Glasgow, Scotland, Aug. 2009.
11. D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788, 1999.
12. C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 1998.
13. D. J. Rohde and O. Cappe. Online maximum-likelihood estimation for latent factor models. In *IEEE conference on statistical signal processing (SSP2011)*, Nice, France, July 2011.
14. R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, volume 20, 2008.